ISSN 1561-5359. Штучний інтелект, 2016, № 2

УДК 68T50

О.О. Марченко, А.О. Никоненко, Т.В. Россада, Є. А. Мельников
Київський національний університет імені Тараса Шевченка, Україна
вул. Володимирська, 64/13, м. Київ, 01601

СИСТЕМА ВИЗНАЧЕННЯ АВТОРСТВА ТЕКСТУ

O.O. Marchenko, A.O. Nykonenko, T.V. Rossada, E.A. Melnikov
Taras Shevchenko National University of Kyiv
Volodymyrska St., 64/13, c. Kyiv, Ukraine, 01601

AUTHORSHIP ATTRIBUTION SYSTEM

Було розроблено систему ідентифікації та перевірки авторства документа, побудовану на основі машинного навчання. Оригінальність моделі обумовлена запропонованим унікальним профілем ознак автора, що дозволив, із застосуванням методу опорних векторів (SVM), отримати високі показники точності.
Ключові слова: ідентифікація авторства, машинне навчання, метод опорних векторів
A new effective system for identification and verification of text authorship has been developed. The system is created on the base of machine learning. The originality of the proposed model is caused by the unique profile of the author attributes that allows getting extra-high performance accuracy using the method of the Support Vector Machine (SVM).
Keywords: authorship identification, machine learning, support vector machine

Вступ

Ідентифікація та перевірка авторства є унікальною і, водночас, дуже затребуваною задачею, з огляду на можливість застосування у різних сферах діяльності людини: для боротьби з плагіатом, для встановлення авторства анонімних текстів, для експертизи та встановлення особистості в криміналістиці та у багатьох інших задачах та напрямах. Задача є також дуже складною через фундаментальну проблему формування набору ознак, за якими можна оцінити ймовірність належності тексту певному автору. Задачу ускладнює також той факт, що до останнього часу для розроблених систем визначення авторства текстів необхідною умовою їх стійкої та якісної роботи була наявність великих об’ємів авторських текстів у навчальній вибірці. Ще однією вадою розроблених моделей є їх якісне обмеження на кількість авторів. Якщо у випадку наявності текстів 3-4 авторів у навчальній та тестовій вибірках навчені класифікатори впевнено демонструють до 85% точності визначення автора тексту у тестовій виборці, то зі зростанням кількості авторів у вибірках до 6 та вище точність класифікації різко падає до 60-55%.

Авторами було розроблено систему визначення авторства текстів за умови наявності мінімального набору текстів для кожного автора у навчальній вибірці (від 5К текстів на кожного автора) та із кількістю авторів до 20.

Система являє собою набір класифікаторів для визначення ідентичності авторського стилю в тексті. На вхід системи подається документ із заявленим авторством і-того письменника. Система перевіряє, наскільки поточні значення ознак вхідного тексту відповідають «еталонним» значенням ознак даного автора. На основі аналізу значень ознак тексту, система підтверджує або спростовує факт приналежності тексту заявленому автору.

Система має дворівневу архітектуру. На першому рівні ряд класифікаторів обчислює оцінку приналежності тексту тому чи іншому автору (з кількості авторів, на яких система навчена). На другому рівні вирішується оптимізаційна задача встановлення єдиного автора тексту, на основі аналізу рішень окремих класифікаторів першого рівня. Якщо, в результаті аналізу рішень, буде встановлено, що текст містить

фрагменти текстів, котрі стилістично не належать заявленому автору, то система робить відповідний висновок.

На стадії навчання класифікатори 1-го рівня навчаються розпізнавати тексти конкретного i-го автора. Навчання проходить на наборі текстів даного автора, в яких присутні 100% ідентичні тексти i-го автора і його ж тексти, розбавлені в різних пропорціях фрагментами інших авторів навчального корпусу (всього корпус містив 15 авторів). Кожен класифікатор вчиться розпізнавати 100% ідентичні тексти свого автора серед текстів з домішками інших авторів. Як базовий алгоритм машинного навчання класифікаторів застосовується лінійний метод опорних векторів (SVM). При цьому використовуються набори багаторівневих ознак-властивостей тексту:

1) ознаки пунктуаційного рівня (статистика розділових знаків);
2) ознаки морфологічного рівня (статистика сполучень літер);
3) ознаки лексичного рівня (N-грами, статистика стоп слів, статистика універсальних слів і т.д.);
4) ознаки синтаксичного рівня (N-грами частин мови, частоти синтаксичних зв'язків і т.д.).

Після того, як кожен класифікатор на відповідному розміченому наборі текстів навчився розпізнавати свого автора, слідує другий етап навчання, що складається в підборі ваг пріоритетів класифікаторів для вирішення колізій, коли одночасно кілька класифікаторів ідентифікують авторство їх письменника.

Після завершення навчання, система функціонує наступним чином. На вхід для перевірки авторського стилю надходить документ з позначеним автором. Документ проходить обробку послідовністю лінгвістичних процесорів, у результаті чого визначаються всі необхідні характеристики вхідного тексту. Це дозволяє системі, на основі отриманих властивостей, обчислити значення для всіх ознак побудованої моделі тексту. Отримані значення ознак використовуються навченими класифікаторами для визначення авторства вхідного тексту. Якщо виникає колізія, і відразу кілька класифікаторів визначають документ як текст свого автора, то ваги пріоритету дозволяють вирішити колізію і визначити єдиного правильного автора. У разі неможливості визначення автора приймається рішення про наявність в тексті запозичених фрагментів.

Характеристики системи

Система призначена для аналізу текстів англійською мовою обсягом від 800 слів, без сленгу та спеціальної лексики, структура тексту має бути подібна до структури новинної статті або есе і не містити великої кількості цитат, діалогів або інших фрагментів специфічної структури. Для кожного тексту вказаний автор (його ID). Система повинна перевірити авторство (за принципом істина/хиба) для кожного документа. Документ складається з авторських фрагментів-абзаців та/або запозичених абзаців-фрагментів. Абзац-фрагмент тексту виділений на початку і кінці символами переходу на новий рядок. Розмір абзацу - від 70 слів. Запозичені у інших авторів фрагменти можуть відрізнятися або збігатися за тематикою з основним текстом. Пропорція авторських фрагментів до запозичень у тексті випадкова (всі фрагменти можуть бути як авторськими, так і запозиченими).

Для попереднього навчання авторському стилю, системі повинна бути надана навчальна вибірка авторських текстів сумарним об'ємом від 12000 слів для кожного автора. Кількість авторів у системі не перевищує 15-20.

Метод

У зв'язку з відносно невеликими об'ємами наявних даних, було прийняте рішення щодо аналізу кожного автора окремо, в подальшому дана модель отримала назву «Document based», тоді як модель, що аналізує кожен текст автора окремо, отримала назву «File based». Далі в цьому розділі розглядаються особливості функціонування кожної з моделей.

Document based версія заснована на припущенні, що для визначення авторського стилю необхідно отримати максимально можливу кількість авторського тексту. Для реалізації даного принципу всі доступні тексти одного автора об'єднуються у документ, який потім аналізується системою. Перевагою даного підходу є отримання більш точних та більш згладжених статистик. Наявність великої кількості даних дозволяє:

1. Гарантовано перейти від ознак, специфічних для конкретного тексту, до ознак, специфічних для автора. Наприклад, авторський текст, написаний для персонального блогу, може стилістично суттєво відрізнятися від тексту автора в газетній або науковій статті. Причому відрізнятися може як лексика, характерні звороти, знаки пунктуації (неформальне листування містить більше знаків питання та оклику), так і структура речень. Побудова гіпотез для машинного навчання на базі різнорідного тексту дозволяє створити універсальні гіпотези, що будуть ефективно працювати на будь-якому типі тексту, за умови, що такий тип входив до документа.
2. Отримати згладжені статистики за рахунок більшої кількості різнорідних даних. Збільшення довірчого інтервалу дозволяє системі виявити авторський стиль у документах з кращою точністю. Будь-яка зібрана статистична інформація (наприклад, середня кількість літер у слові) стає незалежною від стилю тексту і краще відображає авторський шаблон.

File based версія моделі розглядає кожен файл окремо і збирає інформацію для гіпотез. Деякі гіпотези подаються в модуль машинного навчання «as is», інша інформація усереднюється по кожному з авторів для отримання характеристик авторського стилю, близьких по якості до document based моделі**.**

Результатом обробки вхідних даних з навчальної вибірки для document based моделі є унікальний вектор авторських ознак для кожного автора. Результатом обробки вхідних даних для file based моделі є набір векторів по кожному окремому автору. На етапі навчання модель підбирає коефіцієнти для кожної позиції вектора, після чого модель вважається навченою і може класифікувати нові тексти.

Реалізація

З точки зору машинного навчання (Machine Learning, ML), задача визначення авторства зводиться до задачі класифікації тексту - система повинна віднести текст до одного з п'ятнадцяти класів - типова задача мультикласифікації.

Навчання document based моделі проходить на наборі текстів кожного автора, в яких містяться його 100% ідентичні тексти, а також - його тексти, розбавлені в різних пропорціях фрагментами інших авторів з навчального корпусу (всього корпус містить 15 авторів). Кожен класифікатор вчиться розпізнавати 100% ідентичні тексти свого автора від текстів з домішками інших авторів.

Навчання file based моделі проходить на наборах текстів кожного автора, кожен текст 100% належить одному автору, домішки недопустимі. Так як і в попередньому випадку, кожен класифікатор вчиться розпізнавати тексти свого автора.

Як базовий алгоритм машинного навчання класифікаторів, застосовується лінійний метод опорних векторів (linear SVM). Для реалізації моделі використовувалася мова Python та пакети scikit-learn [1] та numpy [2]. Етап виділення

інформативних ознак був найскладнішим у роботі, як базовий набір були реалізовані ознаки (features) зі статей [3-9]. Деякі з цих ознак виявилися неефективними для даної задачі. Наприклад, аналіз помилок, описаний в статті [6]. Не спрацювали і такі ознаки, як word N-grams, описані в [4], такі популярні ознаки, як повнота і об'єм словникового запасу [3]. Виникло припущення, що дані ознаки можуть добре працювати у випадку, коли автори пишуть на одну фіксовану тематику, але не у випадку набору текстів з корпусу новин. Зі схожих причин не спрацювали такі ознаки, як довжина слова/речення, типова перша/остання літера слова/речення, частоти сполучення слів та деякі інші.

Від таких ознак, як dependency triplets [3], k-ee subtree шаблон [8] довелося відмовитися через незначний приріст точності у моделі при великій складності обчислення цих ознак.

До результуючого набору ознак для визначення авторського стилю увійшли:
1) ознаки пунктуаційного рівня (статистика розділових знаків, їх середні значення (means) і стандартне відхилення (standard deviation));
2) ознаки морфологічного рівня (статистика буквосполучень та букв);
3) ознаки лексичного рівня (статистика стоп слів, статистика універсальних слів, авторських слів і т.д.);
4) ознаки для оцінки середньої схожості-зв'язності речень на базі методу[10];
5) ознаки синтаксичного рівня (N-грами частин мови, частоти синтаксичних зв'язків, частоти застосувань правил виведення (з граматик Хомського) і т.д.).
6) складні ознаки на базі векторного представлення статистики використання слів, обчислені на тестовому корпусі.

Після завершення етапу навчання множини класифікаторів (кожен розпізнає свого автора), слідує другий етап навчання, де проводиться підбір ваг коефіцієнтів для вирішення колізій. Колізією є випадок, коли одночасно декілька класифікаторів ідентифікують текст як такий, що належить їх автору. Спеціальний метод OVR (one-vs-rest) вирішує ряд оптимізаційних задач з підбору ваг пріоритетів для класифікаторів, щоб мінімізувати кількість помилок розпізнавання на навчальному наборі текстового корпусу.

Методика випробувань

При виборі методики оцінювання системи, ключову роль зіграли такі фактори, як повторюваність та всебічність оцінки. Метою було створити єдиний корпус, на якому інші дослідники могли перевіряти точність нашої системи, а також виміряти та порівняти результати своїх систем на тому ж тестовому наборі текстів. У період розробки системи, її тестування та оцінка проводилися на авторських текстах статей з The Washington Post, The New York Times, The Daily Telegraph, The Times, The Wall Street Journal та з інших західних англомовних видань. Однак, політика використання даних текстів забороняє їх вільне розповсюдження. Отже, повторення отриманих результатів стає складним у даному випадку.

Після закінчення розробки системи, її перевірка на стабільність проходила на текстах з блогу компанії P1K [11]. Був зібраний корпус з трьох авторів, по десять випадкових текстів кожного автора. На даному корпусі document based версія системи показала досить низькі результати через специфіку її навчання на малій кількості авторів – система мала всього три приклади для навчання (по одному документу, складеному з усіх текстів автора). Очевидно, що така кількість прикладів не є достатньою для machine learning алгоритмів. На цьому етапі була створена file based версія, яка здатна працювати з наборами даних, починаючи від двох унікальних авторів.

Базовим корпусом для оцінювання систем визначення авторства пропонується використовувати RCV1 dataset (Reuters Corpus Volume I) [12]. Для забезпечення повноти оцінки, RCV1 була відсортована за авторами, після чого було виділено шість окремих наборів даних. Для кожного набору даних описано, тексти яких авторів туди були включені і кількість статей кожного автора. Автори відбиралися в алфавітному порядку «as is» і не нормалізувалися за кількістю статей.

1. **Dataset0** – містить 15 авторів, 37 текстів, загальним об'ємом 107.788Б сумарно, тексти кожного автора в RCV1 займають від 4КБ до 9КБ: *A.H. Yoon(2ст., 7160Б); Abdus Sattar(2ст., 4219Б); Adam Cataldo(2ст., 5643Б); Adel Abu Niimeh(3ст., 7609Б); Adrian Blum(2ст., 4917Б); Alan Hoskins(2ст., 6165Б); Aleksandar Mitic(2ст., 5985Б); Alexis Sindahijo(3ст., 9133Б); Alikhan Tasuyev(2ст., 6303Б); Amanda Stultz(3ст., 9992Б); Ana Isabel Martinez(3ст., 9002Б); Anastas Petrov(3ст., 8884Б); Andrea McDaniels(3ст., 5895Б); Andres Rendon(3ст., 9219Б); Andrew Bartram(2ст., 7662Б)*
2. **Dataset1** – містить 15 авторів, 191 текст, загальним об'ємом 542.732Б сумарно, тексти кожного автора в RCV1 займають від 24КБ до 58КБ: *Abdoulaye Massalatchi(13ст., 34697Б); Agnes Tsang(11ст., 36442Б); Alan Dickey(13ст., 25340Б); Alberto Pontes(19ст., 49896Б); Aleksandrs Rozens(13ст., 43649Б); Alfredo Aranda(12ст., 35369Б); Allieu Ibrahim Kamara(12ст., 38178Б); Alma Davanzo(10ст., 32275Б); Amitav Ranjan(12ст., 32091Б); Anderson Fumulani(14ст., 28688Б); Andrius Vilkancas(12ст., 24498Б); Anna Smirnova(11ст., 40750Б); Anna Wardenburg(18ст., 58141Б); Artyom Danielyan(11ст., 36829Б); Ashok Pahalwan(10ст., 25909Б).*
3. **Dataset2** – містить 15 авторів, 848 текстів, загальним об'ємом 2.533.525Б сумарно, тексти кожного автора в RCV1 займають від 130КБ до 197КБ: *Abbas Salman(53ст., 156643Б); Abdelaziz Barrouhi(54ст., 140427Б); Al Yoon(62ст., 190592Б); Alan Elsner(48ст., 162607Б); Alexander Miles(54ст., 139600Б); Ali Bouzerda(74ст., 197417Б); Aline van Duyn(57ст., 192188Б); Alison Leung(42ст., 130917Б); Allan Dowd(50ст., 137048Б); Allan Seccombe(70ст., 195957Б); Andrea Hopkins(47ст., 154429Б); Andrew Gill(53ст., 191142Б); Andrew Steele(50ст., 186356Б); Andrew Stern(50ст., 170870Б); Andy Capostagno(84ст., 187332Б).*
4. **Dataset3** – містить 15 авторів, 1554 тексти, загальним об'ємом 5.107.210Б сумарно, тексти кожного автора в RCV1 займають від 300КБ до 388КБ: *Abigail Levene(91ст., 330243Б); Adam Cox(84ст., 300655Б); Adam Entous(108ст., 330949Б); Adrian Edwards(104ст., 324003Б); Alistair Bell(106ст., 362574Б); Alver Carlson(111ст., 345038Б); Andrew Huddart(97ст., 324710Б); Andrew Kelly(99ст., 331601Б); Andrew Marshall(98ст., 324779Б); Andrew Tarnowski(88ст., 305340Б); Anis Ahmed(99ст., 316438Б); Arthur Malu-Malu(124ст., 372795Б); Bernard Edinger(120ст., 388312Б); Brian Spoors(102ст., 371736Б); Carmel Linnane(123ст., 378037Б).*
5. **Dataset4** – містить 15 авторів, 2440 текстів, загальним об'ємом 8.156.620Б сумарно, тексти кожного автора в RCV1 займають від 504КБ до 593КБ: *Aaron Pressman(187ст., 587900Б); Anatoly Verbin(145ст., 541457Б); Anchalee Koetsawang(179ст., 539921Б); Andrew Gray(164ст., 526620Б); Andrew Hurst(145ст., 545520Б); Anton Ferreira(163ст., 518143Б); Ashraf Fouad(164ст., 519588Б); Ben Hirschler(171ст., 568846Б); Bill Tarrant(142ст., 552036Б); Bradley Perrett(192ст., 582358Б); Brian Williams(153ст., 526783Б); Buchizya Mseteka(174ст., 593874Б); Caroline Brothers(146ст., 541194Б); Chris Bird(155ст., 507869Б); Daniel Sternoff(151ст., 504511Б).*
6. **Dataset5** – містить 15 авторів, 4896 текстів, загальним об'ємом 17.311.692Б сумарно, тексти кожного автора в RCV1 займають від 991КБ до 1,5МБ: *Alan Baldwin(287ст.,*

991888Б); Alastair Macdonald(272ст., 1037306Б); Alexander Smith(320ст., 1183679Б); Alistair Lyon(314ст., 1142824Б); Andrew Hill(298ст., 1018305Б); Carol Giacomo(364ст., 1377804Б); Charles Aldinger(363ст., 1164119Б); Douglas Busvine(290ст., 1032241Б); Ellen Freilich(319ст., 1012316Б); Evelyn Leopold(473ст., 1499613Б); Glenn Somerville(314ст., 1151629Б); Leonard Santorelli(338ст., 1687289Б); Linda Sieg(285ст., 1129334Б); Marcel Michelson(289ст., 998340Б); Martin Cowley(370ст., 1184605Б).

Створення цих шести наборів даних дозволяє провести всебічну оцінку стабільності роботи системи, починаючи від випадку, коли є всього по 2-3 статті кожного автора, до аналізу точності на корпусі з декілька тисяч статей.

Точність будь-якого алгоритму машинного навчання залежить від того, як було поділено дані на навчальний та тестовий набір (train set та test set), тому було б некоректно просто заявити результати на певному наборі даних. Існують методи отримання більш надійних результатів. Наприклад, кросвалідація, але цей метод також залежить від початкового розбиття даних. При зміні параметру seed для функції random результати кросвалідації змінюються, хоч і не так істотно, як у випадку оцінювання з єдиним розбиттям на набори train/test.

Завжди стабільно-однаковий результат на наборі даних дає алгоритм кросвалідації під назвою Leave One Out (LOO). Даний алгоритм не залежить від вибору параметру seed. Суттєвим мінусом LOO є його велика обчислювальна складність: для кожного прикладу з набору даних необхідно провести навчання моделі на всіх інших даних і потім проводити тестування на даному прикладі. Результат, наближений до LOO, дає метод, що має назву simplified LOO.

Simplified LOO виконує перетворення всіх текстів у вектор ознак на самому початку роботи і не обчислює їх заново. Потім з даної матриці ознак послідовно виділяється по одному вектору, який використовується як тестовий приклад, усі інші дані матриці використовуються для навчання моделі. Точність моделі розраховується аналогічно до LOO – як середнє арифметичне усіх прикладів. Simplified LOO допускає можливість певного перенавчання (overfitting), оскільки у побудованій системі існують складні ознаки, що використовують статистики, зібрані з усього корпусу. У випадку LOO в ролі корпусу виступає train set, а у випадку Simplified LOO – весь корпус. Отже, складні ознаки будуть містити також і статистики, зібрані з тестових прикладів. На великих корпусах вплив даного фактору є дуже несуттєвим, а отже, Simplified LOO дає гарне наближення до результатів LOO, проте використовує значно менше обчислювальних ресурсів.

Результати

Для обчислення результатів тестування використовуються три методи: K-fold Crossvalidation, Simplified LOO, LOO. На початку проведення експериментів використовувалося класичне значення K=5 в Crossvalidation, однак, для невеликих наборів даних (Dataset0, Dataset1) при такому малому значенні K частина класів попадає лише в train set, а частина – лише в test set, що суттєво викривлює статистику. Тому вирішено було використовувати значення K, більше за кількість авторів. Значення K=20 не захищає повністю від випадку повного невходження автора в train set, проте дає гарне наближення до цього. Для розбиття наборів даних у K-fold Crossvalidation використовується функція KFold (n_folds=20, shuffle=True, random_state=1) з пакету scikit-learn.

Результати побудованої системи (file based version) можна побачити в таблиці 1. Рядок global містить оцінку F1, обчислену на загальній кількості true positives, false negatives, false positives. Рядок class-based містить оцінку F1, обчислену як середнє арифметичне оцінок F1 кожного класу (кожного автора).

Таблиця 1. Результати роботи системи

		20-Fold Cross Validation	*Cross Validation LOO*	*Simplified LOO*
Dataset0	lobal	Precision=0.7297 Recall=0.7297 F1=0.7297*	Precision=0.7297 Recall=0.7297 F1=0.7297*	Precision=0.9189 Recall=0.9189 F1=0.9189
	lass-based	Precision=0.6789 Recall=0.7222 F1=0.6849*	Precision=0.6844 Recall=0.7222 F1=0.6881*	Precision=0.9333 Recall=0.9222 F1=0.9200
Dataset1	lobal	Precision=0.7068 Recall=0.7068 F1=0.7068	Precision=0.7120 Recall=0.7120 F1=0.7120	Precision=0.8010 Recall=0.8010 F1=0.8010
	lass-based	Precision=0.7218 Recall=0.7014 F1=0.6937	Precision=0.7355 Recall=0.7081 F1=0.6992	Precision=0.8139 Recall=0.7955 F1=0.7918
Dataset2	lobal	Precision=0.7205 Recall=0.7205 F1=0.7205	Precision=0.7252 Recall=0.7252 F1=0.7252	Precision=0.7547 Recall=0.7547 F1=0.7547
	lass-based	Precision=0.7185 Recall=0.7107 F1=0.7062	Precision=0.7238 Recall=0.7168 F1=0.7127	Precision=0.7533 Recall=0.7477 F1=0.7442
Dataset3	lobal	Precision=0.7394 Recall=0.7394 F1=0.7394	Precision=0.7413 Recall=0.7413 F1=0.7413	Precision=0.7606 Recall=0.7606 F1=0.7606
	lass-based	Precision=0.7385 Recall=0.7361 F1=0.7210	Precision=0.7384 Recall=0.7380 F1=0.7223	Precision=0.7596 Recall=0.7580 F1=0.7432
Dataset4	lobal	Precision=0.7434 Recall=0.7434 F1=0.7434		Precision=0.7557 Recall=0.7557 F1=0.7557
	lass-based	Precision=0.7429 Recall=0.7385 F1=0.7305		Precision=0.7556 Recall=0.7511 F1=0.7438
Dataset5	lobal	Precision=0.7680 Recall=0.7680 F1=0.7680		Precision=0.7725 Recall=0.7725 F1=0.7725
	lass-based	Precision=0.7761 Recall=0.7656 F1=0.7619		Precision=0.7805 Recall=0.7702 F1=0.7667

*У даних випробуваннях два класи не розпізналися через малу кількість статей у навчальному наборі даних train set, що суттєво знизило загальну точність системи.

Загалом, як видно з таблиці, система показала феноменально високу точність порівняно з іншими розробками, з огляду на те, що запропонована модель працює з великою кількістю авторів (15-20) і для навчання вимагає відносно малу кількість текстів кожного автора (від 5К).

Висновки

У роботі описується розробка унікальної системи розпізнавання авторства текстів англійською мовою. Система працює на корпусах статей з великим числом авторів (15-20) і демонструє високу точність визначення авторства текстів. При цьому система не вимагає великої навчальної вибірки по кожному автору. Мінімальна вибірка може сягати лише 5К тексту. Використання моделей машинного навчання та розробка унікального профілю ознак авторського стилю дозволили досягнути результату на рівні нового state-of-the-art.

Подяка

Автори статті вдячні компанії P1K і, зокрема, команді проекту Unplug за підтримку в дослідженнях та допомогу в розробці даного алгоритму визначення авторства тексту, в його тестуванні та впровадженні в продукти компанії.

Література

1. Scikit-learn http://scikit-learn.org/stable/
2. Numpy http://www.numpy.org/
3. Fissette, M. Author identification in short texts. Thesis, Department of Artificial Intelligence, 2010, Radboud University.
4. George K. Mikros and Kostas Perifanos Authorship Identification in Large Email Collections: Experiments Using Features that Belong to Different Linguistic Levels - Notebook for PAN at CLEF 2011.
5. Rachel M. Green, John W. Sheppard Comparing Frequency- and Style-Based Features for Twitter Author Identification// Proceedings of the Twenty-Sixth International FLAIRS, St. Pete Beach, Florida, USA, 2013, May 22-24.
6. Roman Kern Grammar Checker Features for Author Identification and Author Profiling// Notebook for PAN at CLEF 2013.
7. Zheng, Rong, Li, Jiexun, Huang, Zan and Chen, Hsinchun. A Framework for Authorship Identification of Online Messages: Writing Style Features and Classification Techniques. Journal of the American Society for Information Science and Technology (JASIST), 57(3):378-393 (2006).
8. Tie-Yun Qian, Bing Liu, Qing Li, Jianfeng Si Review Authorship Attribution in a Similarity Space. Journal of Computer Science and Technology.2015. 30. pp.1200-1213.
9. Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney Authorship attribution using probabilistic context-free grammars. In Proceedings of ACL-2010, pages 38-42.
10. Никоненко А.О., Дослідження статистичної схожості-зв'язності // Вісник КНУ імені Тараса Шевченка, серія фіз..-мат.науки. - 2016. - № 1 - С. 131-136.
11. https://unplag.com/blog/
12. Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research// The Journal of Machine Learning Research, 5, 361-397.

Literatura

1. Scikit-learn http://scikit-learn.org/stable/
2. Numpy http://www.numpy.org/
3. Fissette, M. Author identification in short texts. Thesis, Department of Artificial Intelligence, 2010, Radboud University.
4. George K. Mikros and Kostas Perifanos Authorship Identification in Large Email Collections: Experiments Using Features that Belong to Different Linguistic Levels - Notebook for PAN at CLEF 2011.
5. Rachel M. Green, John W. Sheppard Comparing Frequency- and Style-Based Features for Twitter Author Identification// Proceedings of the Twenty-Sixth International FLAIRS, St. Pete Beach, Florida, USA, 2013, May 22 – 24.
6. Roman Kern Grammar Checker Features for Author Identification and Author Profiling// Notebook for PAN at CLEF 2013.
7. Zheng, Rong, Li, Jiexun, Huang, Zan and Chen, Hsinchun. A Framework for Authorship Identification of Online Messages: Writing Style Features and Classification Techniques. Journal of the American Society for Information Science and Technology (JASIST), 57(3):378-393 (2006).
8. Tie-Yun Qian, Bing Liu, Qing Li, Jianfeng Si Review Authorship Attribution in a Similarity Space. Journal of Computer Science and Technology.2015. 30. pp.1200-1213.

9. Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney Authorship attribution using probabilistic context-free grammars. In Proceedings of ACL-2010, pages 38-42.
10. Nykonenko A.O., Doslidzhennya statystychnoyi skhozhosti-zv"yaznosti // Visnyk KNU imeni Tarasa Shevchenka, seriya fizyko-matematychni nauky. - 2016. - # 1 - C. 131-136.
11. https://unplag.com/blog/
12. Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research// The Journal of Machine Learning Research, 5, 361-397.

RESUME

O.O. Marchenko, A.O. Nykonenko, T.V. Rossada, E.A. Melnikov

Authorship Attribution System

The paper introduces a system that identifies and verifies authorship of the English text document. The system demonstrates high accuracy of the authorship identification and does not require a large training set for each author. The originality of the proposed model is caused by the unique profile of the author attributes that allows getting extra-high performance accuracy using the method of the Support Vector Machine (SVM). The system contains a set of classifiers to determine the identity of the author's style in the text. On the input the system gets a document marked with a label of some author. The system checks how the current values of the text attributes correspond to reference values of the labeled author. On the base of analysis of the input text attributes values the system confirms or denies the fact of the genuine authorship. The system has two-layer architecture. At the first level a set of classifiers calculates the assessments of the probability of belonging the input text to a particular author (authors from the set on which the system is trained). At the second level the optimization problem is solved for determining a single author of the input text by analyzing solutions of the first level classifiers. If the high-level classifier figures out that the text contains pieces that stylistically don't belong to the declared author the system makes an appropriate conclusion. Machine learning uses multi-level features of the text: (1) punctuation features (punctuation statistics); (2) morphological features (letter combinations statistics); (3) lexical features (N-grams, stop-words statistics, etc.); (4) syntactic features (parts of speech N-grams, syntactic dependencies frequency, etc.).

Надійшла до редакції 18.11.2016